CSCI 620/Section 01/Mior, Introduction to Big Data, Fall 2201

Assignment 5 – Data integration

Description

Let's assume the following sources:

- S1: ComedyMovie(id, title, year), which stores comedy movies whose runtime is greater or equal than 75 minutes
- S2: NonComedyMovie(id, title, year), which stores movies whose runtime is greater or equal than 75 minutes and they are not comedies (there is no comedy genre related to them)
- S3: ComedyActor(id, name, birthYear, deathYear), which stores actors who have participated in at least a comedy movie (independent of the runtime)
- S4: NonComedyActor(id, name, birthYear, deathYear), which stores actors who have never participated in any comedy movie (independent of the runtime)
- S5: ActedIn(actor, movie), which stores all actors participation in movies

Let's assume the following global schema:

- All_Movie(id, title, year, genre), which contains each movie with its main genre.
- All_Actor(id, name, birthYear, deathYear), which contains each actor.
- All_Movie_Actor(actor, movie), which stores actors participating in movies.

Your tasks

1. Implement the previous sources in your relational database from assignment 2 using both non-materialized and materialized views. Provide your code. **(20 points)**

2. Using GAV mappings, describe the global schema based solely on the sources. These will be non-materialized views using the views in Q1. You should have two versions of these views using the materialized and non-materialized views from Q1. **(20 points)**

3. Provide queries for answering the following queries over global schema. **(5 points per query)**

    3.1. Alive actors who have participated in more than 10 movies between 2000 and 2005.
    3.2. Actors whose name starts with "Ja" and who have never participated in any comedy movie.

4. Provide the resulting queries after expanding Q3.1 and Q3.2 using the GAV mappings defined in Q2 over the sources. You should simply substitute the definitions of the views into the queries. Run the resulting queries. Report your timings when you use non-materialized vs. materialized views from Q1.
   (25 points)

5. Optimize the queries from Q4 by eliminating redundant joins and sources where possible. Run the resulting queries. Report your timings when you use non-materialized vs. materialized views from Q1.
   (25 points)